JOURNAL OF Evolutionary Biology

doi: 10.1111/jeb.12331

Population genomics of eusocial insects: the costs of a vertebrate-like effective population size

J. ROMIGUIER*, J. LOURENCO*, P. GAYRAL*†, N. FAIVRE*, L. A. WEINERT*‡, S. RAVEL*, M. BALLENGHIEN*, V. CAHAIS*, A. BERNARD*, E. LOIRE*, L. KELLER§ & N. GALTIER*

*Institut des Sciences de l'Evolution de Montpellier, Université Montpellier 2, CNRS UMR 5554, Montpellier, France
†Institut de Recherches sur la Biologie de l'Insecte, CNRS UMR 7261, Université François-Rabelais, Tours, France
‡Department of Veterinary Medicine, University of Cambridge, Cambridge, UK
§Department of Ecology and Evolution, Biophore, University of Lausanne, Lausanne, Switzerland

Keywords:

genomics;
insects;
life-history evolution;
molecular evolution;
population genetics.

Abstract

The evolution of reproductive division of labour and social life in social insects has lead to the emergence of several life-history traits and adaptations typical of larger organisms: social insect colonies can reach masses of several kilograms, they start reproducing only when they are several years old, and can live for decades. These features and the monopolization of reproduction by only one or few individuals in a colony should affect molecular evolution by reducing the effective population size. We tested this prediction by analysing genome-wide patterns of coding sequence polymorphism and divergence in eusocial vs. noneusocial insects based on newly generated RNA-seq data. We report very low amounts of genetic polymorphism and an elevated ratio of nonsynonymous to synonymous changes – a marker of the effective population size – in four distinct species of eusocial insects, which were more similar to vertebrates than to solitary insects regarding molecular evolutionary processes. Moreover, the ratio of nonsynonymous to synonymous substitutions was positively correlated with the level of social complexity across ant species. These results are fully consistent with the hypothesis of a reduced effective population size and an increased genetic load in eusocial insects, indicating that the evolution of social life has important consequences at both the genomic and population levels.

Introduction

One of the most fascinating life-history strategies observed in insects is eusociality, an organization characterized by caste differentiation, division of labour and cooperative brood care, which evolved many times independently in Hymenoptera (wasps, bees, ants), and once in Dictyoptera (termites, Keller & Chapuisat, 2001). Social insect genomics has been developing at a fast rate in recent years (Fischman *et al.*, 2011; Gadau *et al.*, 2012). Functional analyses have yielded interesting insights into the determinants and correlates of caste differentiation and behaviour (e.g. Scharf *et al.*, 2005; Grozinger *et al.*, 2007; Steller *et al.*, 2010; Toth *et al.*, 2010; Kent *et al.*, 2011; Ometto *et al.*, 2011; Wang *et al.*, 2013), and candidate genes/processes associated with the evolution of eusociality (Woodard *et al.*, 2011). Full genome sequence analysis in one bee and seven ant species has also revealed a number of interesting features, such as a low but heterogeneous genomic GC content, a high recombination and gene conversion rate, complete DNA methylation gene sets, depletion in innate immunity genes and enrichment in genes involved in chemical communication (Kent *et al.*, 2012; Gadau *et al.*, 2012; Simola *et al.*, 2013; Libbrecht *et al.*, 2013).

Besides the functional aspects associated with the division of labour, social insect species typically differ from solitary ones in terms of life-history traits. Colonies

Correspondence: Nicolas Galtier, Institut de Sciences de l'Evolution de Montpellier, Université Montpellier 2, CNRS UMR 5554, Place E. Bataillon, 34095 Montpellier, France. Tel.: (+33) 467 14 48 18; fax: (+33) 467 14 36 10; e-mail: nicolas.galtier@univ-montp2.fr

and queens can live up to thirty years (Keller, 1998; Keller & Jemielity, 2006), that is, orders of magnitude longer than most solitary insects. The generation time of eusocial species is therefore considerably lengthened, on average, compared with solitary ones. Another consequence of the peculiar biology of eusocial insects concerns the effective population size (N_e) – a central parameter of the population genetic theory. The very small number of reproductive individuals per colony presumably results in a strong reduction in effective population size in eusocial species due to extreme reproductive skew (Crozier, 1979; Graur, 1985). Colonies of ants, bees and termites can reach a total mass of several kilograms or more (e.g. Shik *et al.*, 2012). One could plausibly speculate that the effective population size of social insect species should be more similar to that of typical vertebrates than to that of typical solitary insects.

The population genetic theory makes several predictions regarding the influence of N_e on molecular variation patterns. First, a reduced N_e is expected to result in decreased amounts of within-species genetic polymorphism – in a panmictic Wright–Fisher population, the expected heterozygosity of a neutral locus is proportional to the product of locus mutation rate by N_e (Kimura, 1983). Secondly, a reduction in N_e is predicted to result in a decreased efficiency of natural selection. N_e is inversely proportional to the rate of genetic drift. Strong genetic drift in small populations is expected to push allele frequencies up and down irrespective of their contribution to fitness, counteracting the effects of natural selection (Kimura, 1983; Ohta, 1987; Lynch, 2007). Assuming that a majority of amino acid changes are deleterious, we would therefore expect a higher ratio of nonsynonymous to synonymous changes in eusocial than in solitary insects, both within species and between species, because of less efficient purifying selection in the former. Interestingly, when they tested this prediction based on 25 eusocial vs. noneusocial contrasts, Bromham and Leys (2005) only obtained equivocal support in favour of the hypothesis of a reduced long-term population size in eusocials. Their study, however, relied on just a handful of genes, which perhaps limited the power of their analysis. Therefore, the prediction of a reduced N_e in eusocial insects due to extreme reproductive skew so far lacks empirical support from DNA sequence data.

Here, we analyse genome-wide patterns of coding sequence polymorphism and divergence in various groups of eusocial insects, which we compare to solitary insects and vertebrates, using both newly generated and publicly available data. We report that the amount of genomic diversity is markedly lower, and the ratio of nonsynonymous to synonymous changes markedly higher, in eusocial than in solitary insect species, in agreement with theoretical expectations. We show that the molecular evolution of protein-coding sequences in eusocial insects is comparable to that of mammals, that is, typical of low-N_e species.

Materials and methods

Sample collection, RNA extraction and sequencing

Five to eleven individuals (imagos) of the sweat bee *Halictus scabiosae* (Hymenoptera, eusocial), harvest ant *Messor barbarus* (Hymenoptera, eusocial), big-headed ant *Pheidole pallidula* (Hymenoptera, eusocial), Glanville fritillary *Melitaea cinxia* (Lepidoptera, solitary), small skipper *Thymelicus sylvestris* (Lepidoptera, solitary) and house mosquito *Culex pipiens* (Diptera, solitary) were collected in 2010, 2011 and 2012 in various localities of their natural geographical range, that is, worldwide in *C. pipiens*, Europe and North Africa in the other species (Table S1). In eusocial species, the sampled individuals were either workers (*H. scabiosae*, *M. barbarus*) or queens (*P. pallidula*). For each individual, whole-body RNA was extracted using the standard protocols as described in Gayral *et al.* (2011), and a non-normalized cDNA library was prepared. The libraries were sequenced on a Genome Analyzer II or Hiseq 2000 (Illumina, Inc.) to produce 100-bp (*H. scabiosae, M. barbarus, P. pallidula, M. cinxia, T. sylvestris*) or 50-bp (*C. pipiens*) single-end fragments (Illumina reads). In addition, for one individual of *H. scabiosae*, a normalized random-primed cDNA library was prepared sequenced for half a run using a 454 Genome Sequencer (GS) FLX Titanium Instrument (Roche Diagnostics). Reads were trimmed of low-quality terminal portions using the SeqClean program (http://compbio.dfci.harvard.edu/tgi/). A similar data set obtained by Gayral *et al.* (2013) in the subterranean termite *Reticulitermes grassei* (Dictyoptera, eusocial) was also included in the analysis.

Transcriptome assembly, read mapping, coding sequence prediction

De novo transcriptome assembly based on the 454 (one individual in *H. scabiosae*) and Illumina reads was performed following strategies B and D in Cahais *et al.* (2012), using a combination of the programs Abyss and Cap3. Reads were mapped to predicted cDNAs (contigs) using the BWA program. In the case of *C. pipiens*, the mapping of the reads was performed on the already available transcriptome of this species downloaded from the site of the Broad Institute (http://www.broadinstitute.org/annotation/genome/culex_pipiens.4). Contigs covered at 25X or less (across all individuals) were discarded. Open reading frames (ORFs) were predicted using the program transcripts_to_best_scoring_ORFs.pl, which is part of the Trinity package. Contigs carrying no ORF longer than 200 bp were discarded.

Calling polymorphic sites and genotypes

At each position of each ORF and each individual, diploid genotypes were called according to the method described by Tsagkogeorga *et al.* (2012, model M1) and improved by Gayral *et al.* (2013), using the reads2snps program. This method first estimates the sequencing error rate in the maximum-likelihood framework, calculates the posterior probability of each possible genotype and retains genotypes supported at >95% – otherwise missing data are called. A minimum of ten reads per position and per individual was required to call a genotype. Positions at which a genotype could be called in less than half the number of available individuals were discarded. Then, single nucleotide polymorphisms (SNPs) were filtered for possible hidden paralogues (duplicated genes) using a likelihood ratio test based on explicit modelling of paralogy (Gayral *et al.*, 2013). Population genomic statistics were calculated using home-made programs that rely on the Bio++ libraries (Guéguen *et al.* 2013). Confidence intervals were obtained by bootstrapping loci. To obtain comparable data in *D. simulans,* we downloaded π_N and π_S estimates for a total of 6702 transcripts from the supplementary data of Begun *et al.* (2007).

The synonymous and nonsynonymous site frequency spectra (i.e. the distribution of minor allele counts across SNPs) were computed based on the predicted genotypes. To cope with the variable sample size across SNPs, a hypergeometric projection of the observed SFS into a subsample of twelve sequences was applied (Hernandez *et al.*, 2007), and SNPs sampled in less than twelve sequences (six diploid individuals) being disregarded. Model parameters were estimated by the maximum-likelihood method following Eyre-Walker *et al.* (2006) using a home-made C program.

Phylogenomic analyses

Annotated coding sequences from the seven fully sequenced ant genomes and the honeybee genome were downloaded from the Fourmidable database (Wurm *et al.*, 2009). 4920 groups of one-to-one orthologues shared by the seven species were predicted using the ORTHO_MCL pipeline (Li *et al.* 2003). Within each group of orthologues, coding sequences were aligned using MACSE (Ranwez *et al.*, 2011). Positions (codon sites) including >75% of missing data were removed. A maximum-likelihood analysis of each alignment was conducted using the codeML program (Yang, 2007) under a model allowing the ratio of nonsynonymous to synonymous substitutions to differ between branches of the tree, using the tree of Gadau *et al.* (2012) as guide tree, with honeybee serving as out-group. Then, for each terminal branch of the tree, the average (across genes) d_N/d_S ratio was calculated, weighting each alignment by its length, after the data set was cleaned for unexpectedly long branches – most of which presumably result from a methodological artefact in case of substitutional saturation. Cleaning was achieved by discarding alignments in which one (or more) of the terminal branch synonymous length was longer than four times the median synonymous length of this branch across the entire data set.

The same analysis was performed in fruit flies (ten species, Drosophila 12 Genomes Consortium, 2007; 9846 groups of predicted orthologues), mosquitoes (four species, Ensembl Metazoa release 20, 4115 orthologues), mammals (twelve species, Orthomam v7, Ranwez *et al.*, 2007, 3913 orthologues) and birds (five species, Ensembl release 73, 8574 orthologues). The list of analysed species is provided in Table S2. In fruit flies, two species, *D. sechelia* and *D. persimilis*, were excluded from the data set because they are very closely related to *D. simulans* and *D. pseudobscura*, respectively, so that their inclusion would create very short branches problematic for substitution rate analysis. In mammals, the twelve species were selected based on criteria of genome coverage and phylogenetic distribution. In each of birds and mosquitoes, all the fully sequenced genomes were used. The phylogenetic trees used in these analyses were obtained from the literature:

[ants] (*Harpegnathos.saltator*, (*Linepithema.humile*, (*Camponotus.floridanus*, (*Pogonomyrmex.barbatus*, (*Solenopsis.invicta*, (*Acromyrmex.echinatior*, *Atta.cephalotes*))))));
[fruit flies] ((*Drosophila.grimshawi*,(*Drosophila.virilis*, *Drosophila.mojavensis*)),(*Drosophila.willistoni*,(*Drosophila.pseudobscura*,(*Drosophila.ananassae*,((*Drosophila.yakuba*, *Drosophila.erecta*),(*Drosophila.simulans*, *Drosophila.melanogaster*))))));
[mosquitoes] ((*Anopheles.gambiae*, *Anopheles.darlingi*), (*Culex.quinquefasciatus*, *Aedes.aegyptii*));
[mammals] (*Loxodonta.africana*,(*Dasypus.novemcinctus*, (((*Sorex.araneus*, *Erinaceus.europaeus*),(*Bos.taurus*,(*Myotis.lucifugus*, *Canis.familiaris*))),(*Homo.sapiens*,(*Oryctolagus.cuniculus*,(*Mus.musculus*, *Dipodomys.ordii*))))));
[birds] ((*Anas.plathyrhynchos*,(*Gallus.gallus*, *Meleagris.gallopavo*)),(*Ficedula.albicolis*, *Taeniopygia.guttata*));

In ants, fruit flies and mammals, the per-site, per-million year average synonymous substitution rate was estimated by first assigning a date to each internal node based on the literature (Hedges *et al.*, 2006; Obbard *et al.*, 2012; Moreau & Bell, 2013), then dividing branch length (measured in per-site synonymous substitutions) by divergence time for each terminal branch and finally taking the average across terminal branches.

Results

Population genomic data sets

The whole-body transcriptome of four to eleven wild-caught individuals from seven species of insects (four

eusocial, three solitary) was Illumina-sequenced. cDNAs, ORFs, SNPs and genotypes were predicted following Gayral *et al.* (2013). In house mosquito, the assembly step did not work properly – a low number of relatively short contigs were obtained, perhaps due to the shorter read length (50 bp) and elevated genetic diversity (see below) in this species. Therefore, we used a publicly available transcriptome assembly based on the house mosquito genome project instead of our *de novo* assembly (see Material and Methods). Population genomic statistics regarding within-species variation in ~8000 genes of the fruit fly *Drosophila simulans* (Diptera), a fourth solitary species, were obtained from Begun *et al.* (2007). Table 1 summarizes the data sets finally obtained and analysed. The number of genes (ORFs) sufficiently covered to be included in the analysis varied from ~3000 house mosquito, sweat bee) to ~13 000 (small skipper), and their average length from 201 (Glanville fritillary) to 1068 (big-headed ant). These numbers are somewhat related to the total amount of data available per species. No significant difference was detected in average number of ORFs and average ORF length between eusocial and solitary species.

The predicted ORFs were compared to the nonredundant NR database using BLASTP. The percentage of ORFs that found no hit (e-value > 0.001) varied between 0.1% (house mosquito) and 13% (subterranean termite). For the ORFs that found at least one hit, we recorded the GenBank taxonomy of the first hit. The percentage of first hits assigned to class Insecta was above 98% in six of the seven analysed species and equal to 89.2% in subterranean termite. In the latter species, the remaining ORFs were assigned in majority to noninsect arthropods (3.4%) or nonarthropod metazoans (6.9%). These numbers presumably reflect discrepancies in available genomic resources – no complete genome has been sequenced in Dictyoptera so far, whereas several are available in each of Hymenoptera, Diptera and Lepidoptera. This analysis suggests that the vast majority, if not the totality, of the ORFs we predicted correspond to genuine insect genes and not contaminants. The results reported below were qualitatively unchanged when we only retained ORFs assigned to class Insecta by first BLAST hit.

Coding sequence polymorphism

Table 2 and Fig. 1 summarize the main population genomic statistics, averaged across genes, in the eight species we analysed. Figure 1a shows a substantially reduced level of within-species genetic diversity, both synonymous (π_S) and nonsynonymous (π_N), in the four eusocial species (closed circles), compared with the four solitary ones (open circles). The average π_S differed by one order of magnitude between the two groups of species (eusocial average: 0.0030; noneusocial average: 0.033; *t*-test, $P = 0.0024$). This is consistent with the hypothesis of a reduced N_e in eusocial species. The mean ratio of nonsynonymous to synonymous polymorphism (π_N/π_S, Fig. 1b) was also higher in eusocial than in solitary insects. The trend was a bit less marked than in Fig. 1a, but still significant (*t*-test, $P = 0.035$). This result is again consistent with the hypothesis of a lower efficiency of purifying selection due to a smaller long-term N_e in eusocial species. Figures 1c,d show the same relationships in log-scale. For the sake of comparison, we added in Table 2 and Fig. 1 two species of mammals for which the same statistics are available from the literature, namely central chimpanzee (*Pan troglodytes troglodytes*, Hvilsom *et al.*, 2012) and Iberian hare (*Lepus granatensis*, Gayral *et al.*, 2013). The four eusocial insect species analysed here appear quite similar to these two mammals regarding coding sequence diversity.

Finally, we investigated the distribution of allele frequency across SNPs, or site frequency spectra (SFS), separating synonymous from nonsynonymous variants (Fig. S1). This was performed in the five species for which nine or more individuals had been analysed. In the three eusocial species, the synonymous SFS were highly similar to the expected distribution of a Wright–Fisher population (i.e. panmictic and constant N_e) under neutrality. The nonsynonymous SFS, however, revealed an excess of low-frequency variants, which is typically interpreted as reflecting the contribution of

Table 1 Population genomic data sets analysed in this study.

Species	No. of individuals	10^6 read (total)	Read length	No. of ORFs	av. ORF length	No. of SNPs
R. grassei (subterranean termite)	9	21.5	50	8412	357	8379
M. barbarus (harvest ant)	10	48.2	100	9393	468	40 605
P. pallidula (big-headed ant)	4	130	100	9121	1068	28 221
H. scabiosae (sweat bee)	11	15.5	100	3287	297	3779
M. cinxia (Glanville fritillary)	10	17.9	100	4272	201	43 631
T. sylvestris (small skipper)	7	115	100	13262	492	171 934
C. pipiens (house mosquito)	10	21.5	50	3290	333	84 967
D. simulans	7	NA	NA	7978	590	NA

Table 2 Genome-wide synonymous and nonsynonymous diversity in two mammals, four eusocial and four solitary insect species.

Group	Species	π_S (%)	π_N (%)	π_N/π_S	References*
Mammal	*P. troglodytes* (chimpanzee)	0.20	0.046	0.22	(1)
Mammal	*L. granatensis* (Iberian hare)	0.41 [0.38–0.44]	0.060 [0.05–0.07]	0.15 [0.13–0.17]	(2)
Termite	*R. grassei* (subterranean termite)	0.11 [0.11–0.12]	0.023 [0.019–0.025]	0.21 [0.19–0.22]	(2)
Ant	*M. barbarus* (harvest ant)	0.58 [0.57–0.60]	0.053 [0.051–0.055]	0.091 [0.087–0.095]	(3)
Ant	*P. pallidula* (big-headed ant)	0.31 [0.30–0.32]	0.053 [0.051–0.055]	0.17 [0.16–0.18]	(3)
Bee	*H. scabiosae* (sweat bee)	0.21 [0.20–0.23]	0.026 [0.023–0.029]	0.12 [0.11–0.14]	(3)
Butterfly	*M. cinxia* (Glanville fritillary)	3.4 [3.3–3.5]	0.32 [0.30–0.34]	0.094 [0.088–0.099]	(3)
Butterfly	*T. sylvestris* (small skipper)	2.3 [2.2–2.3]	0.15 [0.14–0.15]	0.065 [0.063–0.067]	(3)
Mosquito	*C. pipiens* (house mosquito)	4.1 [4.0–4.2]	0.11 [0.10–0.12]	0.027 [0.025–0.029]	(3)
Fruit fly	*D. simulans*	3.3	0.22	0.066	(4)

*References: (1) Hvilsom *et al.* (2012); (2) Gayral *et al.* (2013); (3) this study; (4) Begun *et al.* (2007).

slightly deleterious mutations to polymorphism (Keightley & Eyre-Walker, 2007). The situation was a bit different in house mosquito and Glanville fritillary, in which the synonymous SFS departed the neutral Wright–Fisher expectation, implying a more complex demographic/migratory history for these species. The nonsynonymous SFS showed an excess of low-frequency variants in these two species too, even though the difference with the neutral SFS appeared less marked – but still highly significant (multinomial test, $P < 10^{-10}$ in all five species).

For each of these five species, we fit a population genetic model in which synonymous mutations are assumed to be neutral and nonsynonymous mutations deleterious. The population selection coefficient of nonsynonymous mutations was assumed to be distributed according to a negative gamma distribution of shape parameter beta and mean $S = 4N_e s$, following Eyre-Walker *et al.* (2006). Model parameters were estimated in the maximum-likelihood framework (Table S3). The estimated average strength of negative selective, S, was of the order of 10^3 in eusocial species and much higher in solitary species (10^6–10^9), consistent again with the hypothesis of a lower N_e in the former. The estimated shape parameter of the estimated distribution of fitness effect of nonsynonymous mutations was between 0.16 and 0.37, that is, similar to values similarly obtained in *Homo sapiens* and *Drosophila melanogaster* (Keightley &

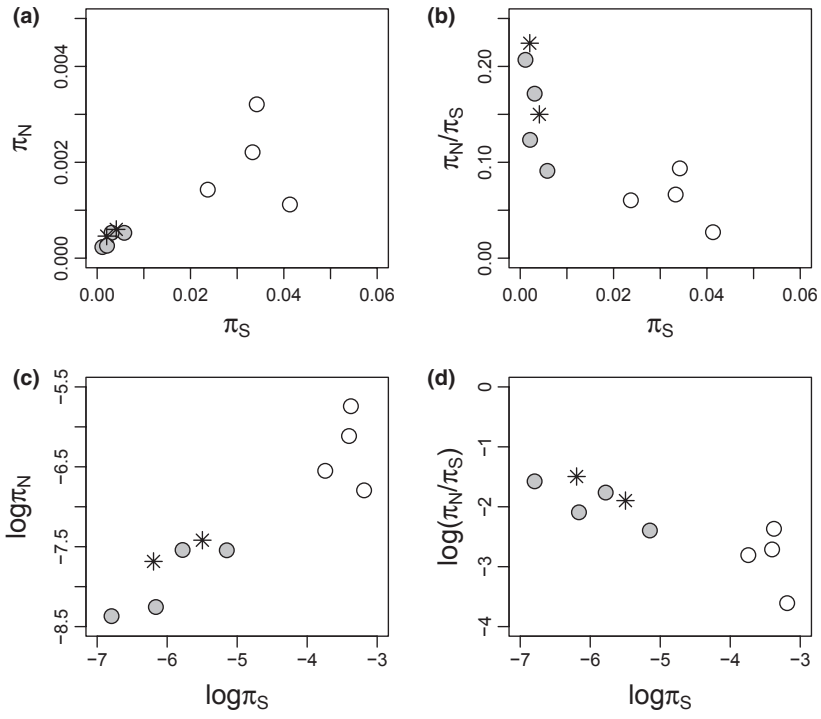

Fig. 1 Population genomic pattern in eusocial vs. solitary insect species. Closed circles: eusocial species. Open circles: solitary species. Stars correspond to two mammalian species: hare and chimpanzee. Data from Table 2.

Eyre-Walker, 2007). Figure 2 shows the estimated proportion of nearly neutral, slightly deleterious, mildly deleterious and strongly deleterious nonsynonymous mutations in the five analysed species. This figure confirms that the rate of slightly deleterious mutations (relative to effective population size) is substantially higher in the three low-N_e eusocial species than in the two high-N_e solitary ones, which results in a heavier mutation load in the former.

Phylogenomic analyses

Multispecies data sets of aligned coding sequences were gathered from public genomic databases in one group of social insects (ants), two groups of solitary insects (fruit flies and mosquitoes) and two groups of vertebrates (mammals and birds). These data sets were analysed in a phylogenetic maximum-likelihood framework. In each group, an estimate of the average (across genes) ratio of nonsynonymous (d_N) to synonymous (d_S) substitutions was obtained for each terminal branch of the tree (Fig. 3, Table S2).

In fruit flies (ten species), the lineage-specific average d_N/d_S ratio varied between 0.05 and 0.1, in agreement with published results (Begun *et al.*, 2007; Heger & Ponting, 2007). In mammals (twelve species), our results were consistent with the previously reported positive correlation between d_N/d_S and both body mass and longevity (Nikolaev *et al.*, 2007; Romiguier *et al.*, 2013), with, for example, the mouse and the shrew showing relatively low d_N/d_S, whereas the elephant and the long-lived bat (maximal longevity: 40 years) showed the highest values. Importantly, the lowest d_N/d_S estimate in mammals was higher than the highest d_N/d_S in fruit flies, offering an appropriate comparative framework for eusocial insects. Interestingly, the estimated d_N/d_S in ants essentially spanned the same range of values as in mammals, that is, between 0.1 and 0.18. The ant average d_N/d_S ratio, 0.131, was not significantly different from the mammalian value (0.146), but significantly higher ($P < 0.001$) than the one in fruit flies (0.074).

The results from mosquitoes (4 species) and birds (5 species) essentially corroborated these contrasts, despite a smaller sample size in these two groups. The five species of birds analysed in this study (chicken, turkey, duck, zebra finch and flycatcher) showed a mean d_N/d_S ratio markedly higher than that of mammals, which was somewhat unexpected. Analysing 8384 orthologues of chicken and zebra finch and using Anolis (lizard) and mammals as out-group, Nam *et al.* (2010) reported d_N/d_S ratios of 0.12-0.15 in avian lineages, that is, similar to our estimates in mammals. The reasons for the discrepancy between the two studies are unclear.

We checked that substitutional saturation was not an issue by examining synonymous branch lengths. We found that the largest d_N/d_S values were associated with short branches (e.g. *Homo sapiens* in mammals, *Atta cephalotes* in ants), whereas the lowest d_N/d_S estimates were obtained from long branches (e.g. *D. willistoni* in fruit flies, *Harpegnathos saltator* in ants). The opposite effect (i.e. underestimated d_S leading to an overestimated

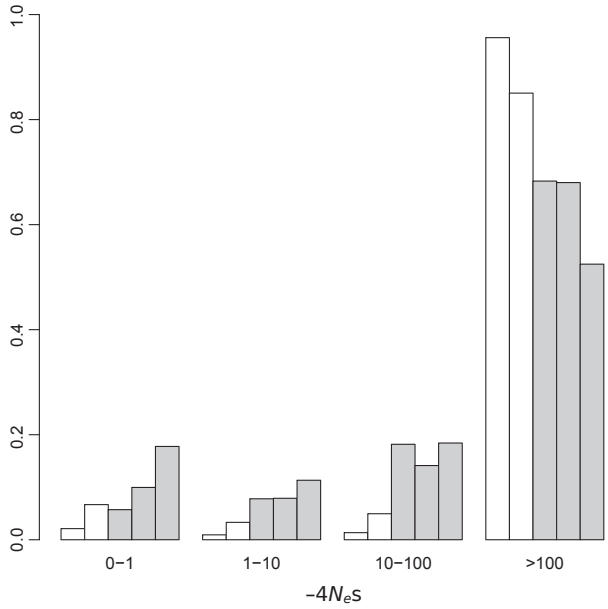

Fig. 2 Estimated distribution of fitness effects of deleterious nonsynonymous mutations in three eusocial and two solitary insect species. *X*-axis: intervals of -4Nes. *Y*-axis: relative contribution of each interval to the distribution. From left to right: *C. pipiens, M. cinxia, M. barbarus, H. scabiosae, R. grassei*. Eusocial species are coloured in grey.

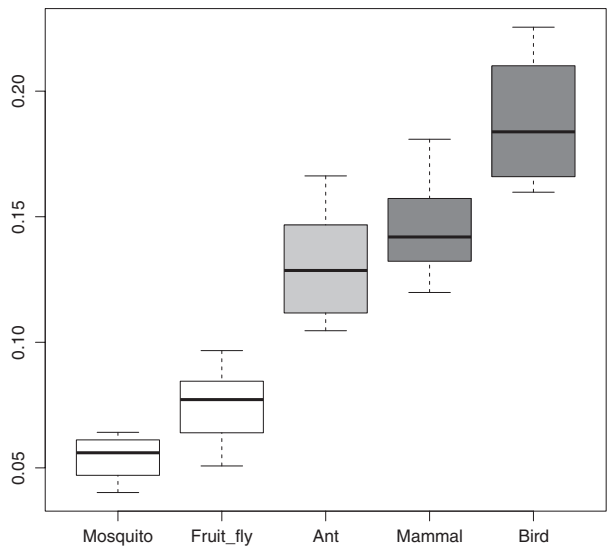

Fig. 3 Distribution of the genome-wide, lineage-specific d_N/d_S ratio in five groups of animals.

d_N/d_S in fast-evolving lineages) would be expected in case of substitutional saturation. When we turned synonymous branch lengths into per-million year substitution rates using published divergence dates, we found that ants evolve at a rate of 0.0044 synonymous substitution per synonymous site per My, which is roughly twice as fast as mammals, but three times slower than fruit flies, on average. Note that the neutral substitution rate, unlike the neutral within-species diversity, is expected to be equal to the mutation rate and independent of N_e (Kimura, 1983).

The seven sequenced ant species are quite diverse in terms of life-history traits (Gadau *et al.*, 2012). Across these species, we detected a significantly positive correlation between d_N/d_S and the queen/worker size ratio ($P = 0.014$), a variable typically used as a measure of the level of differentiation between castes and degree of social complexity (Fig. 4). The jumping ant *H. saltator* showed the lowest d_N/d_S estimate among all seven species. This species, which belongs to the Ponerinae subfamily, forms small colonies of morphologically similar individuals with all physiologically capable of laying eggs. It is the 'least social' of the panel in terms of colony size and caste differentiation. At the other extreme, the leaf-cutting, fungi-growing *Atta cephalotes*, which forms extraordinarily large, long-lived colonies with highly differentiated castes, shows the highest d_N/d_S value, similar to estimates obtained in large-sized, long-lived mammals such as elephant. The relationship between d_N/d_S and the queen/worker size ratio was no longer significant when phylogenetic nonindependence between species was accounted for (phylogenetic contrast analysis, $P = 0.14$). None of the other traits analysed in this study (Table S4) did correlate significantly with d_N/d_S, including queen size and queen longevity.

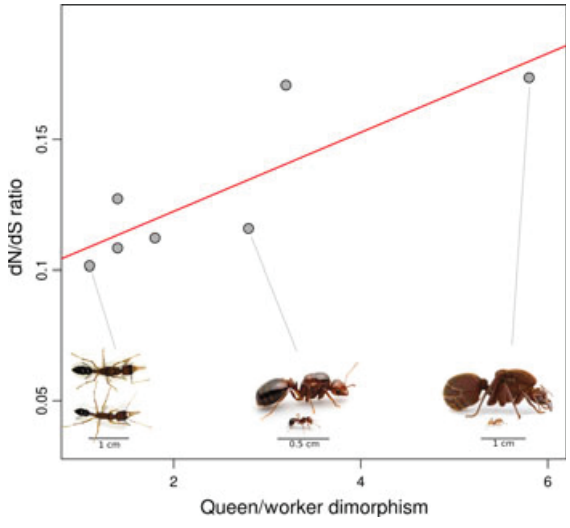

Fig. 4 Correlation between d_N/d_S and queen/worker dimorphism in ants. Queen/worker dimorphism is defined as the ratio of head width between the queen and the smallest cast of worker; illustration in *Harpegnathos saltator*, *Solenopsis Invicta* and *Atta cephalotes*, from left to right. Ant pictures: © Alex Wild, used with permission.

Discussion

In this study, we compared the population genomics of four eusocial and four solitary insects based on transcriptome next-generation sequencing data. We were able to assemble predicted cDNAs and ORFs, confirm they correspond to genuine insect coding sequences, and call SNPs and genotypes for thousands of expressed genes using the approach recently validated by Gayral *et al.* (2013). We investigated patterns of coding sequence variation within species based on this data set and between species using published genomes of ants, fruit flies, mosquitoes, mammals and birds. Two predictions regarding the influence of N_e on molecular evolution were tested, that is, a reduced amount of genetic diversity and a higher ratio of nonsynonymous to synonymous changes in eusocial species.

Reduced genetic diversity in eusocial insects

We report that the four eusocial species of this study show a markedly reduced genome-wide level of coding sequence genetic diversity, both synonymous and nonsynonymous, compared with solitary insects. This is consistent with the hypothesis of a lower N_e in eusocial than in noneusocial insects, in agreement with the theoretical expectation associated with the division of reproductive labour, by which just a tiny fraction of individuals contribute to reproduction. The contrast is strong: genome-wide π_S and π_N differ by one order of magnitude between the two groups of species. As far as genetic diversity is concerned, our four species of eusocial insects are more similar to mammals than to solitary insects. Besides the hare and the chimpanzee shown in Fig. 1, Perry *et al.* (2012) recently analysed the genetic diversity of 16 mammalian species and reported a genome-wide average π_S that varied between 0.1% and 0.6%, that is, the very order of magnitude of the values obtained for eusocial insects in this study.

Besides N_e, the per-generation mutation rate is another important determinant of species genetic diversity (e.g. Nabholz *et al.*, 2008). Even though direct measurements are lacking, it seems plausible to speculate that this rate is higher in eusocial than in solitary insects due to the extended lifespan of the former, which presumably results in an increased number of germ cell divisions per generation (see Lynch, 2010). According to our molecular clock, the neutral substitution rate in ants is only three times as slow as in fruit flies, although their generation time (years) is orders of magnitude longer than that of fruit flies (weeks), corroborating the hypothesis of a higher per-generation

mutation rate in eusocial insects. If this was actually the case, then the contrast in N_e between eusocial and solitary species would be even sharper than the contrast in genetic diversity we are reporting.

The strongest reduction in N_e in this study is observed in the termite *R. grassei*, which shows the lowest π_N, the lowest π_S and the highest π_N/π_S of all the analysed species. We note that in termite colonies, reproduction is monopolized by a single female (queen) and a single male (king), whereas polyandry and polygyny are common in ants (Bourke & Franks, 1995). We speculate that strict monogamy might explain an additional reduction in N_e in termites, compared with eusocial hymenopterans. Genetic data from additional termite species will of course be needed to test this prediction.

An excess of nonsynonymous changes in eusocial insects

Our results are also indicative of an increased π_N/π_S ratio due to less efficient purifying selection in eusocial species. The effect seems rather strong in subterranean termites, in which the genome average π_N/π_S ratio is high and similar to chimpanzees, but not so marked in sweat bees and harvest ants, which did not differ much from solitary insects. In hymenopterans, males are haploid, so that recessive deleterious mutations are exposed to selection in males, which increases the efficiency of purging (Pamilo *et al.*, 1978; Hedrick & Parker, 1997; Glémin, 2007) and might reduce the π_N/π_S. However, our species sample size is clearly too small to conclude on this aspect, and the elevated π_N/π_S (0.17) obtained in big-headed ants does not seem confirm this suggestion. Besides the average π_N/π_S, we also report in sweat bees, harvest ants and subterranean termites a significant difference in allele frequency distribution between nonsynonymous and synonymous SNPs – that is, a substantial excess of low-frequency nonsynonymous variants. This is again similar to the patterns reported in several mammals (Fay *et al.*, 2001; Gayral *et al.*, 2013) and suggestive of a heavy load of deleterious mutations segregating in populations of eusocial insects.

These results are corroborated by our between species analysis of fully sequenced ant genomes. The d_N/d_S ratio we report in various ant lineages is similar to that of mammals and significantly higher than in fruit flies and mosquitoes. Comparing the closely related termite species *R. grassei* and *R. flavipes* (4761 ORFs), Gayral *et al.* (2013) consistently reported an elevated genome average d_N/d_S ratio (0.26) in this genus of eusocial insects. These results are again indicative of a much smaller long-term N_e in eusocial than in solitary insects. Finally, we found a positive relationship between the d_N/d_S ratio and the queen/worker dimorphism, which is a marker of between-caste differentiation, but not between d_N/d_S and queen size or queen longevity, suggesting that colony life-history traits rather than merely phenotypic characteristics of reproductive individuals influence the strength of genetic drift and the efficiency of natural selection in ants.

We note that, besides N_e, the increased fixation probability of nonsynonymous mutations in ants could be explained by GC-biased gene conversion, a recombination-associated repair bias that has been shown to strongly impact genomic landscapes in mammals, birds (Duret & Galtier, 2009), bees and ants (Kent *et al.*, 2012), but much less weakly so in Drosophila (Haddrill & Charlesworth, 2008; Robinson *et al.*, 2014). Galtier *et al.* (2009) demonstrated that episodic GC-biased gene conversion might result in substantial increases in the d_N/d_S ratio. Kent *et al.* (2012) reported in honeybee a faster rate of evolution of GC-enriched genes, compared with the genome average. Interestingly, this effect seems to be associated with cast differentiation: in honeybee, genes that are over-expressed in workers compared with queens tend to be GC-enriched and fast-evolving (Kent *et al.*, 2012).

The high nonsynonymous/synonymous ratio reported in, for example, leaf-cutting ants and subterranean termites reveals the existence of a heavy load of slightly deleterious mutations, both segregating and fixed, which might reflect the high level of specialization and advanced social organization of these species. The elevated genetic load we report from coding sequence analysis presumably affects other aspects of genomic evolution, such as gene content and conservation of regulatory elements, as suggested by Simola *et al.* (2013). Clearly, care should be taken to account for the confounding effect of a reduction in N_e when it comes to interpreting specificities of social insect molecular evolutionary patterns.

Phylogenetic inertia?

Our panel of insect species includes one representative of Dictyoptera, three Hymenoptera (eusocial), two Lepidoptera and two Diptera (solitary). This is obviously a limited sample of the insect diversity. It could be that the contrast we are revealing between eusocial and noneusocial species is in part explained by specificities of these four orders. Strictly speaking, none of the results we report here, taken separately, is significant at the 5% level when method correcting for phylogenetic inertia, such as phylogenetic contrasts, is applied. If, for instance, one calculates one π_S value per order by taking the average across species, then the difference between eusocial (Hymenoptera average: 0.37%; Dictyoptera average: 0.11%) and solitary (Lepidoptera average: 2.85%; Diptera average: 3.70%) orders is only marginally significant (*t*-test, $P = 0.068$).

For several reasons, however, it seems unlikely that the relationships we report are entirely explained by

phylogenetic effects. First, we note that N_e has a very low phylogenetic inertia, especially at the time scale of this study, casting some doubts on the necessity and relevance of the contrast analysis (Lynch, 2011). Secondly, we report several lines of evidence from independent data sets – polymorphism, divergence, between orders, within ants – all supportive of the hypothesis of a smaller N_e in eusocial taxa. That this hypothesis, which corresponds to the theoretical expectation, was favoured by chance several times independently in the absence of an effect would appear implausible. For instance, just combining the phylogeny-corrected *P*-values obtained with the between-order π_S analysis (0.068) and the within-ant d_N/d_S analysis (0.14) using the *Z*-transform test (Whitlock, 2005) yields a meta-analysis *P*-value of 0.035.

Conclusions

This study reveals low levels of genetic diversity and an increased nonsynonymous/synonymous ratio in eusocial insects compared with solitary ones, consistent with the hypothesis of a reduced N_e in eusocials. This is, to our knowledge, the first study unambiguously supporting this intuitive hypothesis. In this case, the biology and life history of species appear to consistently affect their effective population size, overcoming the contingent effects of species demographic history, such as bottlenecks, population expansion and population structure. We note that this result was only revealed when we moved to the genomic scale, suggesting that the true relationships between species life-history traits and genetic diversity could generally be stronger than suggested by currently available data sets (Leffler *et al.*, 2012).

Acknowledgments

We are highly grateful to I. Hanski, S. Ikonen, J. Kullberg, Z. Kolev, M. Weill, C. Atyame Nten, M. Chapuisat, N. Brand, Y. Chiari, G. Tsagkogeorga, A. Lenoir and R. Blatrix for their help with sampling. We thank the Montpellier Bioinformatics & Biodiversity platform for support regarding computational aspects. This work was supported by European Research Council advanced Grant 232971 (PopPhyl) and Agence Nationale de la Recherche Grant ANR-10-BINF-01-01 (Ancestrome) to NG, an advanced European Research Council grant and several grants of the Swiss National Science Foundation to LK.

References

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.P., Hahn, M.W. *et al.* 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* **5**: e3e10.

Bourke, A.F.G. & Franks, N.R. 1995. *Social Evolution in Ants*. Princeton University Press, Princeton, New Jersey, USA.

Bromham, L. & Leys, R. 2005. Sociality and the rate of molecular evolution. *Mol. Biol. Evol.* **22**: 1393–1402.

Cahais, V., Gayral, P., Tsagkogeorga, G., Melo-Ferreira, J., Ballenghien, M., Weinert, L. *et al.* 2012. Reference-free transcriptome assembly in non-model animals from next generation sequencing data. *Mol. Ecol. Resour.* **12**: 834–845.

Crozier, R. 1979. Genetics of sociality. In: *Social Insects* (H.R. Hermann, ed), pp. 223–286. Academic Press, New York.

Drosophila 12 Genomes Consortium 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**: 203–218.

Duret, L. & Galtier, N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* **10**: 285–311.

Eyre-Walker, A., Woolfit, M. & Phelps, T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* **173**: 891–900.

Fay, J.C., Wyckoff, G.J. & Wu, C.I. 2001. Positive and negative selection on the human genome. *Genetics* **158**: 1227–1234.

Fischman, B.J., Woodard, S.H. & Robinson, G.E. 2011. Molecular evolutionary analyses of insect societies. *Proc. Natl. Acad. Sci. USA* **108**: 10847–10854.

Gadau, J., Helmkampf, M., Nygaard, S., Roux, J., Simola, D.F., Smith, C.R. *et al.* 2012. The genomic impact of 100 million years of social evolution in seven ant species. *Trends Genet.* **28**: 14–21.

Galtier, N., Duret, L., Glémin, S. & Ranwez, V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino-acid changes in primates. *Trends Genet.* **25**: 1–5.

Gayral, P., Weinert, L., Chiari, Y., Tsagkogeorga, G., Ballenghien, M. & Galtier, N. 2011. Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. *Mol. Ecol. Resour.* **11**: 650–661.

Gayral, P., Melo-Ferreira, J., Glémin, S., Bierne, N., Carneiro, M., Nabholz, B. *et al.* 2013. Reference-free population genomics from Next-Generation transcriptome data and the vertebrate-invertebrate gap. *PLoS Genet.* **9**: e10003457.

Glémin, S. 2007. Mating systems and the efficacy of selection at the molecular level. *Genetics* **177**: 905–916.

Graur, D. 1985. Gene diversity in hymenoptera. *Evolution* **39**: 190–199.

Grozinger, C.M., Fan, Y., Hoover, S.E. & Winston, M.L. 2007. Genome-wide analysis reveals differences in brain gene expression patterns associated with caste and reproductive status in honey bees (*Apis mellifera*). *Mol. Ecol.* **16**: 4837–4848.

Guéguen, L., Gaillard, S., Boussau, B., Gouy, M., Groussin, M., Rochette, N.C. *et al.* 2013. Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol.* **30**: 1745–1750.

Haddrill, P.R. & Charlesworth, B. 2008. Non-neutral processes drive the nucleotide composition of non-coding sequences in Drosophila. *Biol. Lett.* **4**: 438–441.

Hedges, S.B., Dudley, J. & Kumar, S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics* **22**: 2971–2972.

Hedrick, P.W. & Parker, J.D. 1997. Evolutionary genetics and genetic variation of haplodiploids and X-linked genes. *Annu. Rev. Ecol. Syst.* **28**: 55–83.

Heger, A. & Ponting, C.P. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. *Genome Res.* **17**: 1837–1849.

Hernandez, R.D., Williamson, S.H., Zhu, L. & Bustamante, C.D. 2007. Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. *Mol. Biol. Evol.* **24**: 2196–2202.

Hvilsom, C., Qian, Y., Bataillon, T., Li, Y., Mailund, T., Sallé, B. *et al.* 2012. Extensive X-linked adaptive evolution in central chimpanzees. *Proc. Natl. Acad. Sci. USA* **109**: 2054–2059.

Keightley, P.D. & Eyre-Walker, A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**: 2251–2261.

Keller, L. 1998. Queen lifespan and colony characteristics in ants and termites. *Ins. Soc.* **45**: 235–246.

Keller, L. & Chapuisat, M. 2001. Eusociality and Co-operation. In:*Nature Encyclopedia of Life Sciences*. Nature Publishing Group, London.

Keller, L. & Jemielity, S. 2006. Social insects as a model to study the molecular basis of ageing. *Exp. Gerontol.* **41**: 553–556.

Kent, C.F., Issa, A., Bunting, A.C. & Zayed, A. 2011. Adaptive evolution of a key gene affecting queen and worker traits in the honey bee, *Apis mellifera*. *Mol. Ecol.* **20**: 5226–5235.

Kent, C.F., Minaei, S., Harpur, B.A. & Zayed, A. 2012. Recombination is associated with the evolution of genome structure and worker behavior in honey bees. *Proc. Natl. Acad. Sci. USA* **109**: 18012–18017.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Leffler, E.M., Bullaughey, K., Matute, D.R., Meyer, W.K., Ségurel, L., Venkat, A. *et al.* 2012. Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* **10**: e1001388.

Li, L., Stoeckert, C.J. Jr & Roos, D.S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**: 2178–2189.

Libbrecht, R., Oxley, P.R., Kronauer, D.J. & Keller, L. 2013. Ant genomics sheds light on the molecular regulation of social organization. *Genome Biol.* **14**: 212.

Lynch, M. 2007. *The Origins of Genome Architecture*. Sinauer Associates Inc, Sunderland.

Lynch, M. 2010. Evolution of the mutation rate. *Trends Genet.* **26**: 345–352.

Lynch, M. 2011. Statistical inference on the mechanisms of genome evolution. *PLoS Genet.* **7**: e1001389.

Moreau, C.S. & Bell, C.D. 2013. Testing the museum versus cradle biological diversity hypothesis: phylogeny, diversification, and ancestral biogeographic range evolution of the ants. *Evolution* **67**: 2240–2257.

Nabholz, B., Mauffrey, J.F., Bazin, E., Galtier, N. & Glémin, S. 2008. Determination of mitochondrial genetic diversity in mammals. *Genetics* **178**: 352–361.

Nam, K., Mugal, C., Nabholz, B., Schielzeth, H., Wolf, J.B., Backström, N. *et al.* 2010. Molecular evolution of genes in avian genomes. *Genome Biol.* **11**: R68.

Nikolaev, S.I., Montoya-Burgos, J.I., Popadin, K., Parand, L., Margulies, E.H., NIH Intramural Sequencing Center Comparative Sequencing Program *et al.* 2007. Life-history traits drive the evolutionary rates of mammalian coding and noncoding genomic elements. *Proc. Natl. Acad. Sci. USA* **104**: 20443–20448.

Obbard, D.J., Maclennan, J., Kim, K.W., Rambaut, A., O'Grady, P.M. & Jiggins, F.M. 2012. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Mol. Biol. Evol.* **29**: 3459–3473.

Ohta, T. 1987. Very slightly deleterious mutations and the molecular clock. *J. Mol. Evol.* **26**: 1–6.

Ometto, L., Shoemaker, D., Ross, K.G. & Keller, L. 2011. Evolution of gene expression in fire ants: the effects of developmental stage, caste, and species. *Mol. Biol. Evol.* **28**: 1381–1392.

Pamilo, P., Varvio-Aho, S.L. & Pekkarinen, A. 1978. Low enzyme gene variability in Hymenoptera as a consequence of haplodiploidy. *Hereditas* **1**: 93–99.

Perry, G.H., Melsted, P., Marioni, J.C., Wang, Y., Bainer, R., Pickrell, J.K. *et al.* 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Res.* **22**: 602–610.

Ranwez, V., Delsuc, F., Ranwez, S., Belkhir, K., Tilak, M.K. & Douzery, E.J. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evol. Biol.* **7**: 241.

Ranwez, V., Harispe, S., Delsuc, F. & Douzery, E.J. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLoS ONE* **6**: e22594.

Robinson, M.C., Stone, E.A. & Singh, N.D. 2014. Population genomic analysis reveals no evidence for GC-biased gene conversion in *Drosophila melanogaster*. *Mol. Biol. Evol.* **31**: 425–433.

Romiguier, J., Ranwez, V., Douzery, E.J.P. & Galtier, N. 2013. Genomic evidence for large, long-lived ancestors to placental mammals. *Mol. Biol. Evol.* **30**: 5–13.

Scharf, M.E., Wu-Scharf, D., Zhou, X., Pittendrigh, B.R. & Bennett, G.W. 2005. Gene expression profiles among immature and adult reproductive castes of the termite *Reticulitermes flavipes*. *Insect Mol. Biol.* **14**: 31–44.

Shik, J.Z., Hou, C., Kay, A., Kaspari, M. & Gillooly, J.F. 2012. Towards a general life-history model of the superorganism: predicting the survival, growth and reproduction of ant societies. *Biol. Lett.* **8**: 1059–1062.

Simola, D.F., Wissler, L., Donahue, G., Waterhouse, R.M., Helmkampf, M., Roux, J. *et al.* 2013. Social insect genomes exhibit dramatic evolution in gene composition and regulation while preserving regulatory features linked to sociality. *Genome Res.* **23**: 1235–1247.

Steller, M.M., Kambhampati, S. & Caragea, D. 2010. Comparative analysis of expressed sequence tags from three castes and two life stages of the termite *Reticulitermes flavipes*. *BMC Genomics* **11**: 463.

Toth, A.L., Varala, K., Henshaw, M.T., Rodriguez-Zas, S.L., Hudson, M.E. & Robinson, G.E. 2010. Brain transcriptomic analysis in paper wasps identifies genes associated with behaviour across social insect lineages. *Proc. Biol. Sci.* **277**: 2139–2148.

Tsagkogeorga, G., Cahais, V. & Galtier, N. 2012. The population genomics of a fast evolver: high levels of diversity, functional constraint and molecular adaptation in the tunicate *Ciona intestinalis*. *Genome Biol. Evol.* **4**: 740–749.

Wang, J., Wurm, Y., Nipitwattanaphon, M., Riba-Grognuz, O., Huang, Y.C., Shoemaker, D. *et al.* 2013. A Y-like social chromosome causes alternative colony organization in fire ants. *Nature* **493**: 664–668.

Whitlock, M.C. 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* **18**: 1368–1373.

Woodard, S.H., Fischman, B.J., Venkat, A., Hudson, M.E., Varala, K., Cameron, S.A. *et al.* 2011. Genes involved in convergent evolution of eusociality in bees. *Proc. Natl. Acad. Sci. USA* **108**: 7472–7477.

Wurm, Y., Uva, P., Ricci, F., Wang, J., Jemielity, S., Iseli, C. *et al.* 2009. Fourmidable: a database for ant genomics. *BMC Genomics* **10**: 5.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**: 1586–1591.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Figure S1 Nonsynonymous vs. synonymous site frequency spectra.
Table S1 Origin of the analysed samples.
Table S2 Lineage-specific genome average dN/dS ratio in five groups of animals.
Table S3 Site Frequency Spectrum maximum-likelihood analysis.
Table S4 Life-history traits of the seven analysed ant species.

Received 21 November 2013; revised 27 December 2013; accepted 2 January 2014